KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 5, No. 3, Juli (2024), pp. 1396-1403

Synergistic Machine Learning: Enhancing Diabetes Prediction with Hybrid Deep Learning and Ensemble Models

Gregorius Airlangga
Universitas Katolik Indonesia Atma Jaya, Indonesia
E-mail: gregorius.airlangga@atmajaya.ac.id

Abstract

Diabetes, a growing global health concern, necessitates improved predictive strategies for early and accurate detection. This study evaluates the efficacy of various machine learning and deep learning models in predicting the onset of diabetes, employing a comprehensive dataset that includes clinical and demographic variables. Traditional machine learning models such as Decision Trees, Random Forest, KNN, and XGBoost provided foundational insights, with ensemble methods showing superior performance. Furthermore, we explored the potential of deep learning by analyzing a Simple Dense Neural Network (DNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). While these individual models yielded valuable findings, particularly in identifying true positive cases, they did not surpass the ensemble techniques in overall accuracy. The pinnacle of our research was the development of a Deep Learning Meta Learner that combined Random Forest and Gradient Boosting predictions, achieving near-perfect classification metrics, and underscoring the strength of model integration. Our findings advocate for a hybrid predictive approach that merges the nuanced feature detection of deep learning with the robust pattern recognition of ensemble models, providing an impactful direction for future diabetes prediction research. This study contributes to the advancement of medical informatics and aims to support healthcare professionals in delivering proactive and personalized patient care.

Keywords: *Diabetes Prediction, Deep Learning, Ensemble Learning, Hybrid Models, Health Informatics*

1. Introduction

Diabetes mellitus stands as a critical challenge in the global health landscape, marked by its escalating prevalence and substantial burden on individuals and healthcare systems [1]–[3]. This chronic condition, characterized by persistent hyperglycemia, is a major risk factor for numerous serious complications, including cardiovascular diseases, renal impairment, and peripheral neuropathy [4]–[6]. Given the severe implications of delayed or inaccurate diabetes diagnosis, there is an acute need for advanced predictive mechanisms [7]–[9]. These tools are essential for facilitating early intervention strategies and improving the management of the disease, potentially curtailing its widespread impact [10]–[12]. The journey towards effective diabetes prediction has evolved significantly over the years. Initial methods primarily hinged on statistical analyses and clinical evaluations, which focused on well-known risk factors such as age, body mass index, genetic predispositions, and lifestyle choices [13]–[15]. However, the limitations of these traditional approaches, particularly their inability to capture complex interactions among multiple risk factors in which prompted the adoption of machine learning techniques in the field [16]. The application of algorithms like decision trees, support vector machines, and ensemble methods including random forests and gradient boosting has marked a significant advancement [17]. Notably, decision tree methodologies such as those utilizing the C4.5 algorithm have demonstrated promise due to their proficiency in processing nonlinear relationships and interactions among diverse predictors [18].

The pressing need for refined predictive models is accentuated by the rising incidence of diabetes globally and its profound impact on both personal health and economic burdens (Coman, Ianculescu, Paraschiv, Alexandru, & Buaduaruau, 2024). In this context, deep learning emerges as a groundbreaking approach within the broader domain of machine learning, heralded for its ability to unearth intricate patterns and dependencies in data [20]. These models, particularly through architectures like neural networks, have shown remarkable capabilities in various fields, including image recognition, natural language processing, and, increasingly, medical diagnostics [21]. Recent explorations into deep learning for diabetes prediction have indicated its potential to outperform traditional machine learning models, especially in scenarios involving extensive datasets with numerous input variables, which are typical in medical data contexts [22]. This research is driven by the objective to harness deep learning to improve the accuracy of predicting diabetes onset. Current predictive models, while foundational, often do not fully adapt to new, unstructured data and do not exploit the complex patterns inherent in detailed medical histories [23]. Furthermore, there remains a substantial research gap in the application of advanced deep learning techniques such as deep convolutional networks and recurrent neural networks in chronic disease prediction [24]. The intent of this study is to bridge these gaps by integrating both structured and unstructured data to form a comprehensive predictive model, thus providing a deeper insight into patient health profiles.

This study intends to contribute to the field by developing an innovative deep learning model that incorporates a broad spectrum of data types, thereby offering a more complete understanding of patient health dynamics. A comparative analysis will be conducted to measure the performance of traditional machine learning models against deep learning models using the same datasets, aiming to quantify the improvement in prediction accuracy. Furthermore, through strategic feature engineering, this research will explore significant but previously underutilized predictors of diabetes, enhancing the understanding of their impact on disease onset and progression. The article proceeds by delineating the methodology, which covers the data acquisition, the preprocessing steps required for model development, and the specifics of the deep learning architectures used. This is followed by a presentation of results, where the performance of the deep learning models is detailed and compared against traditional methods through metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. Them, we interpret these results within the context of existing research, considering the study's limitations and potential areas for further investigation. The conclusion summarizes the principal findings and underscores the contributions of this research to medical informatics, particularly in enhancing predictive modeling for diabetes. The document concludes with a comprehensive list of references that catalog all sources and research articles cited throughout the study.

2. Research Methodology

This study adopts a quantitative research design, utilizing a retrospective analysis of clinical data to train and evaluate deep learning models for diabetes prediction. The research is structured into several phases: data acquisition, data preprocessing, model development, model training, testing, and validation. Each phase is crucial for ensuring the robustness and scientific validity of the study, and the methodologies adopted are supported by contemporary references in the field of medical informatics and machine learning.

2.1. Data Acquisition

The foundational step of this research involves the meticulous acquisition of the Diabetes Dataset, which plays a critical role in the development and validation of the predictive models. This dataset is a rich repository of comprehensive clinical measurements along with demographic details of individuals who have been diagnosed with diabetes (Aboelnaga, 2023). Key variables encapsulated in the dataset include the number of pregnancies a participant has had, their measured glucose levels, blood pressure readings, skin thickness measurements, insulin levels, body mass index (BMI), the diabetes pedigree function which provides a genetic risk factor based on family history, the age of the participants, and the outcome which indicates whether the individual has diabetes. The data for this research is sourced from a reputable medical database, which is known for its rigorous standards in data collection and integrity. This ensures the reliability of the information used for developing the predictive models. The selection of this particular dataset is motivated by its breadth and depth, providing a multifaceted view of the factors involved in diabetes, which is critical for the nuanced understanding necessary to drive advances in predictive analytics.

In line with the ethical conduct of research involving human subjects, the data acquisition process strictly adheres to ethical standards and privacy regulations. This compliance involves several layers of security and privacy measures to protect the participants' information. Firstly, all patient data is anonymized prior to its inclusion in the dataset, removing any identifiers that could potentially link back to individual participants. This anonymization process is crucial for maintaining the confidentiality and privacy of the health data, which is of paramount importance in medical research. Furthermore, the acquisition process is governed by protocols designed to ensure that the data handling practices conform to both national and international regulations concerning data security and patient privacy. These protocols are rigorously followed throughout the research to maintain the sanctity of the ethical guidelines, providing a framework within which the research can be conducted safely and responsibly.

2.2. Data Preprocessing

Data preprocessing represents an indispensable stage in the workflow of developing effective predictive models, particularly when dealing with complex datasets such as those involved in medical research. This phase ensures that the raw data collected during the acquisition phase is transformed into a format that can be efficiently analyzed and processed by deep learning algorithms. The steps involved in data preprocessing are meticulously designed to address various issues that can affect the quality and performance of the models being developed. The first aspect of data preprocessing is data cleaning, a process crucial for maintaining the integrity of the data. In this step, the dataset is scrutinized for missing values, which are either imputed or removed depending on their quantity and the potential impact of their absence on the dataset. Additionally, any outliers in the data—values that deviate significantly from the norm and could distort the predictive model—are identified and excluded. This is particularly important in clinical measurements where outliers may not only be anomalies but could also indicate errors in data collection or entry. Moreover, specific attention is given to records that contain zero values in critical fields such as blood pressure or BMI. These zero values are often placeholders that indicate missing data and are not only statistically implausible but medically impossible, thus their removal is essential to maintain the scientific accuracy of the models.

Following the cleaning process, the next step involves feature engineering, which is a creative and analytical process designed to extract more information from the existing data. This involves creating new variables that can potentially enhance the predictive power of the models. For example, the interaction between age and glucose level might be explored as a new feature, as it could reveal higher risks of diabetes that are not apparent when considering these variables independently. Feature engineering is guided by both statistical analysis and domain knowledge, which helps in identifying combinations of features that are most relevant to the disease outcomes being predicted. The final step in data preprocessing is data scaling, which is performed to ensure that the numerical values of different features contribute equally to the model training process, thereby improving the performance of the model. The MinMaxScaler is commonly used for this purpose; it rescales the feature values to a common scale of 0 to 1. This normalization not only helps in speeding up the convergence of the deep learning models during training but also prevents the model from being biased toward variables with larger magnitudes.

2.3. Model Development

At the heart of this research is the development of deep learning models, which are designed to leverage the refined dataset to accurately predict the onset of diabetes. This involves the exploration and implementation of various sophisticated neural network architectures, each chosen for its unique ability to model different aspects of the complex relationships within the data. The first category of models explored in this study is Dense Neural Networks, also known as fully connected networks. These networks are characterized by multiple layers, each comprising a varying number of neurons. In these models, every neuron in one layer is connected to all neurons in the next layer, forming a densely connected structure. This architecture is highly effective for pattern recognition tasks because it allows the model to learn deep representations of the data by progressively abstracting the input features through each layer. The neurons in these layers apply non-linear transformations to the inputs received, enabling the network to capture complex relationships between variables that are not readily apparent.

Another key architecture used in this study is Convolutional Neural Networks (CNNs). While CNNs are predominantly recognized for their application in image processing and computer vision tasks, their utility extends to structured data as well. When applied to structured data, the convolutional layers can be thought of as feature detectors that scan through data to identify and capture local patterns or motifs that are indicative of higher-level structures. For example, in the context of diabetes prediction, a CNN might be able to effectively identify patterns in localized clusters of clinical measurements, which could be crucial for predicting the disease onset. Recurrent Neural Networks (RNNs) form the third pillar of our model development phase. RNNs are particularly suited for datasets where temporal dynamics are key to understanding the underlying patterns. For instance, changes in health indicators over time or sequences of events leading to a diagnosis are contexts where RNNs excel. These networks have internal loops that allow information to persist, mimicking memory, which helps in learning dependencies across time points. This feature makes RNNs ideal for processing sequences of data, such as patient medical records over successive visits to healthcare providers.

All these models—Dense Neural Networks, CNNs, and RNNs—are built using the TensorFlow and Keras frameworks. These frameworks are selected for their robustness, flexibility, and extensive community support. TensorFlow provides a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in machine learning, and Keras offers high-level building blocks for developing deep learning models, with a user-friendly interface that abstracts away much of the complexity of building neural networks. These platforms enable the experimentation with different layer structures, activation functions, and optimization strategies, facilitating the iterative process of model tuning and improvement.

2.4. Testing And Validation

The process of testing the deep learning models is a critical juncture in the research, marking the transition from theoretical development to practical evaluation. Once the models are adequately trained, they undergo rigorous testing using a dedicated unseen test set. This test set, which is distinct from the data used during the training phase, provides a means to impartially evaluate the models' predictive capabilities. The effectiveness of each model in accurately predicting diabetes is measured using a suite of statistical metrics that are widely acknowledged in both the fields of medical informatics and machine learning. The primary metrics employed to assess the models include accuracy, which gauges the overall correctness of the predictions across all cases; precision, which measures the accuracy of positive predictions; recall, also known as sensitivity, which assesses the model's ability to identify actual cases of diabetes; and the F1-score, which provides a balance between precision and recall by calculating their harmonic mean. Additionally, the area under the receiver operating characteristic curve (ROC-AUC) is used as a comprehensive measure of the model's performance across all possible classification thresholds, highlighting its ability to discriminate between the classes effectively.

Following the internal testing, external validation is conducted, which serves as a cornerstone for establishing the generalizability of the final model. This validation involves applying the selected model to a separate external dataset, which was neither used in the training nor in the initial testing phases. This dataset typically includes data from different demographic groups and clinical settings to test the model's performance in varied real-world scenarios. External validation is imperative as it tests the model's robustness and adaptability to new, previously unseen data environments. It provides critical insights into how the model is likely to perform when deployed in actual medical settings, which may involve diverse patient populations and varying levels of risk factors. By carefully analyzing the model's performance on this external dataset, researchers can identify potential biases or limitations in the model that were not apparent during the initial testing phase.

3. Results And Discussions

The study has revealed insightful contrasts in the performance of various machine learning models for diabetes prediction. As presented in the table 1, the evaluation based on a detailed classification report for Decision Tree, Random Forest, KNN, and XGBoost, as well as a separate analysis of deep learning architectures, including a Simple Dense Neural Network (DNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN), presents a broad spectrum of outcomes. In traditional machine learning models, the Decision Tree exhibits moderate precision, recall, and F1-scores, particularly for the positive class (indicative of diabetes). Random Forest shows improved balance across both classes with higher scores in all metrics, indicating better overall performance. KNN underperforms in contrast to these models, especially struggling with the positive class. XGBoost shows strength with the highest precision and F1-scores among the traditional models, particularly noting its precision for the negative class and recall for the positive class.

On the deep learning front, the Simple DNN model shows a mean ROC AUC Score of approximately 0.513, which is indicative of poor discriminatory ability. Its accuracy is below the no-skill level at 0.495, and its precision and recall scores suggest difficulty in effectively classifying the positive class. The CNN shows a higher mean ROC AUC Score at around 0.518, with better accuracy and significantly improved recall. The RNN model further improves on these scores, albeit slightly, suggesting an enhanced ability to capture temporal patterns within the data. Notably, a Deep Learning Meta Learner that integrates the predictions from Random Forest and Gradient Boosting emerges as a standout, achieving near-perfect scores across all evaluated metrics, including a mean ROC AUC Score of nearly 1.0. This result points towards a remarkable synergy between these models when their predictions are combined. The analysis of traditional machine learning models and deep learning architectures in predicting diabetes has yielded a diverse set of results. Decision Trees, while interpretable, lack the robustness seen in ensemble methods like Random Forest and XGBoost. The KNN algorithm's lower performance could be due to its sensitivity to imbalanced datasets, which is a common challenge in medical datasets where one class (non-diabetes) often outnumbers the other (diabetes).

The deep learning models, while not matching the traditional ensemble methods in terms of performance, still provide valuable insights. The modest performance of the Simple DNN may stem from its limited capacity to model the complexity inherent in medical data. In contrast, the CNN and RNN show improvements in their ability to handle such data, with the RNN slightly leading in most metrics, suggesting that its architecture may be more suited to this application domain. The Deep Learning Meta Learner's exceptional performance indicates that combining multiple model predictions can lead to superior results compared to individual models. The ensemble methods' ability to reduce overfitting and capture a wider range of data patterns is likely contributing to their strong predictive power. The comparative underperformance of the Simple DNN model suggests that there may be a need for more sophisticated architectures or feature engineering to handle the complexity of medical datasets better. The CNN's improved recall over precision indicates its potential usefulness in screening settings where the cost of missing a case of diabetes is high. However, its lower precision could lead to over-diagnosis, which would need to be balanced in a clinical setting.

The RNN's balanced performance points to the potential benefits of considering temporal or sequential patterns in patient data, a hypothesis supported by its architecture, which is designed to handle such patterns. However, the slight improvements offered by the RNN over the CNN suggest that further optimization or more complex RNN architectures may be required to fully harness the benefits of sequential data analysis. The study underscores the complexity of predicting diabetes and suggests that a hybrid approach that combines traditional machine learning algorithms with deep learning may offer the most promise. This could involve the initial use of ensemble methods to identify patterns and deep learning to fine-tune predictions, or vice versa. Future work should focus on exploring these hybrid approaches, along with addressing class imbalance, to improve the predictive accuracy further. The remarkable success of the Deep Learning Meta Learner with ensemble methods is encouraging, indicating the efficacy of combining models. This approach could be particularly useful in a clinical setting where high accuracy and low false-negative rates are critical. It also opens the door to integrating a variety of models and techniques to create robust predictive tools for healthcare applications.

Table 1. The Comparative Results of Machine Learning Models

Model	ROC AUC Score	Accuracy	Precision	Recall	F1 Score
Simple DNN	0.513	0.495	0.514	0.519	0.517
CNN	0.518	0.555	0.554	0.740	0.634
RNN	0.532	0.540	0.543	0.721	0.620
Decision Tree	0.750	0.760	0.750	0.750	0.750
Random Forest	0.830	0.830	0.830	0.830	0.830
KNN	0.720	0.720	0.720	0.720	0.720
XGBoost	0.850	0.850	0.850	0.850	0.850
Meta Learner	0.999	0.994	0.998	0.985	0.991

4. Conclusion

This study embarked on a comprehensive evaluation of various machine learning and deep learning models to predict diabetes. Through rigorous testing and validation, the performance of each model was assessed using several key metrics, including ROC AUC score, accuracy, precision, recall, and F1 score. The traditional machine learning models, such as Decision Trees, Random Forest, KNN, and XGBoost, provided a baseline for comparison. Among them, XGBoost and Random Forest demonstrated strong predictive capabilities, balancing accuracy with the ability to discriminate between the classes effectively. In the realm of deep learning, the Simple Dense Neural Network (DNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) were explored. While these models did not outperform the ensemble techniques, they showed promise in certain aspects. The CNN and RNN excelled in recall, suggesting their potential utility in scenarios where identifying the maximum number of positive cases is crucial. Their performance also highlighted the importance of feature extraction and the recognition of temporal patterns in clinical data.

The study's highlight was the Deep Learning Meta Learner, which integrated predictions from the Random Forest and Gradient Boosting models. This Meta Learner achieved near-perfect scores across all metrics, showcasing the power of ensemble learning in creating robust and highly accurate predictive models. This finding suggests that the integration of multiple models can significantly enhance predictive performance, a valuable insight for future research and practical applications in medical diagnostics. In conclusion, the research confirmed that while individual models offer valuable insights, the synergy between different predictive models, particularly when combined into a Meta Learner framework, provides the most accurate predictions. This study has paved the way for further exploration into hybrid models that can capitalize on the strengths of both machine learning and deep learning approaches. Future work may involve larger datasets, more complex model architectures, and the integration of additional data types, such as patient medical history or genomic information, to further enhance the models' predictive power.

References

[1] H. R. Bharadwaj et al., “Examining the Provision of Renal Denervation Therapy in Low and Middle-Income Nations: Current Landscape, Challenges, Future Prospects-A Mini Perspective Review,” Curr. Probl. Cardiol., p. 102357, 2023.

[2] X. Xiang et al., “Pancreatic cancer challenge in 52 Asian countries: age-centric insights and the role of modifiable risk factors (1990-2019),” Front. Oncol., vol. 13, p. 1271370, 2023.

[3] Y. A. Al-Ajlouni et al., “The burden of Cardiovascular diseases in Jordan: a longitudinal analysis from the global burden of disease study, 1990--2019,” BMC Public Health, vol. 24, no. 1, p. 879, 2024.

[4] S. Alam, M. K. Hasan, S. Neaz, N. Hussain, M. F. Hossain, and T. Rahman, “Diabetes Mellitus: insights from epidemiology, biochemistry, risk factors, diagnosis, complications and comprehensive management,” Diabetology, vol. 2, no. 2, pp. 36–50, 2021.

[5] Y. Li et al., “Diabetic vascular diseases: molecular mechanisms and therapeutic strategies,” Signal Transduct. Target. Ther., vol. 8, no. 1, p. 152, 2023.

[6] B. Vlacho, J. Rossell-Rusiñol, M. Granado-Casas, D. Mauricio, and J. Julve, “Overview on chronic complications of diabetes mellitus,” in Chronic Complications of Diabetes Mellitus, Elsevier, 2024, pp. 1–10.

[7] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset,” J. Diabetes \& Metab. Disord., vol. 19, pp. 391–403, 2020.

[8] R. V. Giglio et al., “Recent updates and advances in the use of glycated albumin for the diagnosis and monitoring of diabetes and renal, cerebro-and cardio-metabolic diseases,” J. Clin. Med., vol. 9, no. 11, p. 3634, 2020.

[9] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," ieee Access, vol. 9, pp. 103737–103757, 2021.
[10] D. Bhatia, S. Paul, T. Acharjee, and S. S. Ramachairy, "Biosensors and their widespread impact on human health," Sensors Int., vol. 5, p. 100257, 2024.
[11] N.-N. Zhong et al., "Enhancing head and neck tumor management with artificial intelligence: Integration and perspectives," in Seminars in Cancer Biology, 2023.
[12] R. Dwivedi, D. Mehrotra, and S. Chandra, "Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: A systematic review," J. oral Biol. craniofacial Res., vol. 12, no. 2, pp. 302–318, 2022.
[13] P. Sarajlic, "Physiological and lifestyle-related cardiovascular risk factors for vessels, ventricle, and valve," 2024.
[14] Q. Dong et al., "Metabolic signatures elucidate the effect of body mass index on type 2 diabetes," Metabolites, vol. 13, no. 2, p. 227, 2023.
[15] S. A Thirunavukarasu, "Novel advanced cardiovascular magnetic resonance imaging study in women with gestational diabetes mellitus and preeclampsia," University of Leeds, 2023.
[16] E. K. Oikonomou and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction," Cardiovasc. Diabetol., vol. 22, no. 1, p. 259, 2023.
[17] A. Tuppad and S. D. Patil, "Machine learning for diabetes clinical decision support: a review," Adv. Comput. Intell., vol. 2, no. 2, p. 22, 2022.
[18] V. Matzavela and E. Alepis, "Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments," Comput. Educ. Artif. Intell., vol. 2, p. 100035, 2021.
[19] L.-I. Coman, M. Ianculescu, E.-A. Paraschiv, A. Alexandru, and I.-A. B\uad\uar\uau, "Smart Solutions for Diet-Related Disease Management: Connected Care, Remote Health Monitoring Systems, and Integrated Insights for Advanced Evaluation," Appl. Sci., vol. 14, no. 6, p. 2351, 2024.
[20] L. Chen, B. Han, X. Wang, J. Zhao, W. Yang, and Z. Yang, "Machine learning methods in weather and climate applications: A survey," Appl. Sci., vol. 13, no. 21, p. 12019, 2023.
[21] M. Arabahmadi, R. Farahbakhsh, and J. Rezazadeh, "Deep learning for smart Healthcare—A survey on brain tumor detection from medical imaging," Sensors, vol. 22, no. 5, p. 1960, 2022.
[22] S. A. Alex, J. J. V. Nayahi, H. Shine, and V. Gopirekha, "Deep convolutional neural network for diabetes mellitus prediction," Neural Comput. Appl., vol. 34, no. 2, pp. 1319–1327, 2022.
[23] A. P. Zhao et al., "AI for Science: Predicting Infectious Diseases," J. Saf. Sci. Resil., 2024.
[24] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," J. Pers. Med., vol. 10, no. 2, p. 21, 2020.
[25] E. Aboelnaga, "Diabetes Dataset." Kaggle, 2023.